Phân Tích Dữ Liệu Thực Tế với Python

Bài 5.2: Phân Tích Mô Tả

Quang-Khai Tran, Ph.D
CyberLab, 03/2023

(Ảnh: Internet)

CYBERLEARN
ĐÀO TẠO CHUYÊN GIA LẬP TRÌNH

CYBERSOFT
ĐÀO TẠO CHUYÊN GIA LẬP TRÌNH

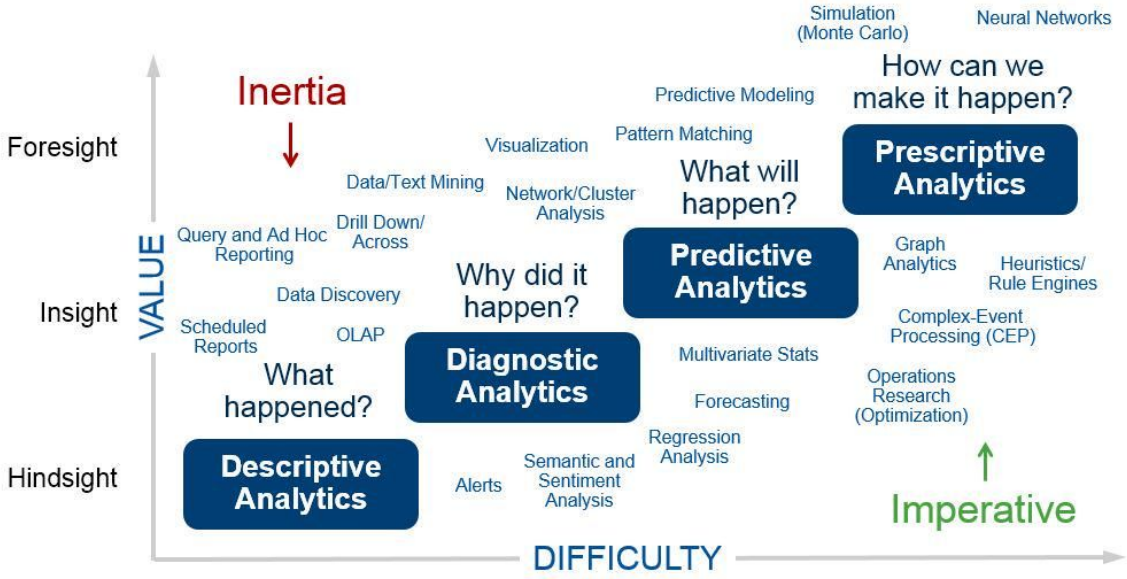
CYBERLAB
ĐÀO TẠO CHUYÊN GIA LẬP TRÌNH

Nội dung

1. Giới thiệu
2. Một số phân tích mô tả đơn biến cơ bản
3. Một số phân tích mô tả đa biến cơ bản
4. Bài tập & Thảo Luận

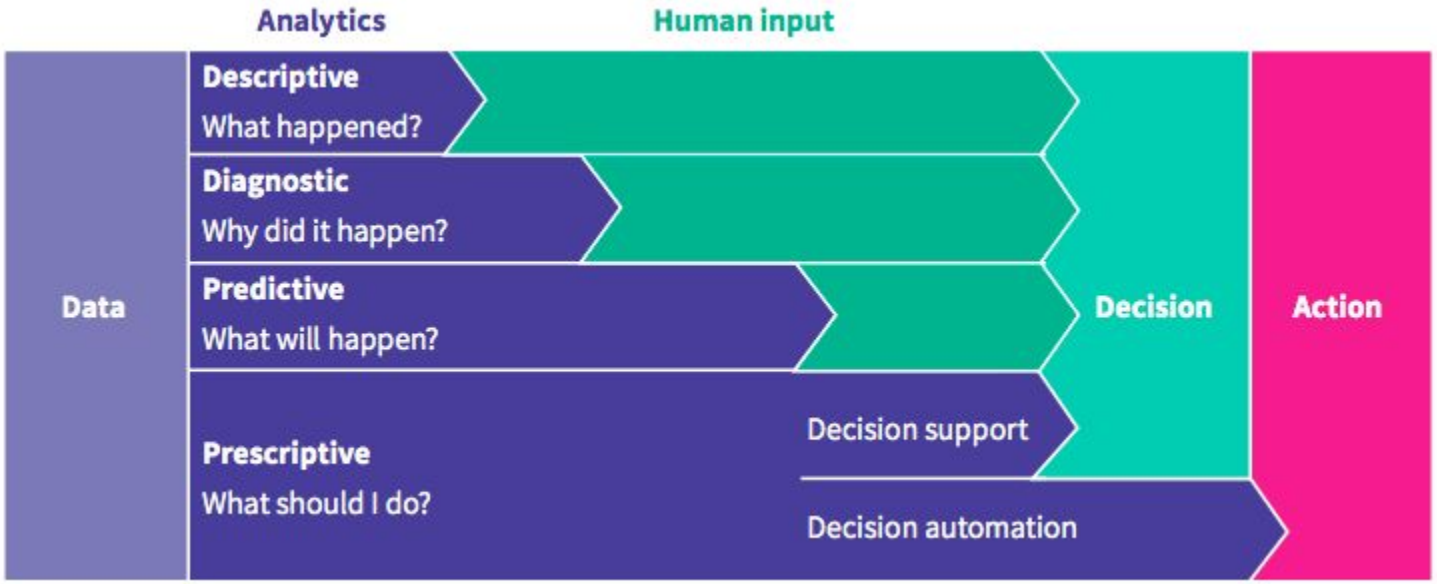
Phần 1. Giới thiệu

The Gartner Analytic Continuum



Gartner

1 Giới thiệu


Analytics
Human input
Data
Descriptive
What happened?
Diagnostic
Why did it happen?
Predictive
What will happen?
Prescriptive
What should I do?
Decision support
Decision automation
Decision
Action

1.2 Phân tích mô tả là gì?

- Là việc tìm hiểu dữ liệu hoặc nội dung (thường là không tự động)
- Trả lời câu hỏi “Điều gì đã diễn ra?” (hoặc Điều gì đang diễn ra?)
- Được thực hiện bởi:
  - Các mô hình BI truyền thống
  - Các công cụ visualization: pie, bar, line …
  - Các bảng và bản tường thuật/câu chuyện được sinh ra (generated narratives)

Descriptive Analytics is the examination of data or content, usually manually performed, to answer the question “**What happened**?” (or **What is happening**?), characterized by traditional business intelligence (BI) and visualizations such as pie charts, bar charts, line graphs, tables, or generated narratives.
(Gartner Glossary: https://www.gartner.com/en/information-technology/glossary/descriptive-analytics)

1.2 Phân tích mô tả là gì?

Phân tích mô tả vs. Thống kê mô tả:

- Có hai hình thức thống kê chính:
  - Thống kê mô tả (descriptive statistics):
    nhằm vào mô tả bản chất của tập dữ liệu thực tế mà bạn đang có
  - Thống kê suy luận (inferential statistics)
    dựa vào dữ liệu đang có để suy ra những đặc tính của tổng thể

⇒ Phân tích mô tả có thể bao gồm cả 2

1 Phân Nhóm Các Loại dữ liệu trong thống kê

Quantitative (định lượng):

- discrete (rời rạc)
- continuous (liên tục)
- interval (khoảng)

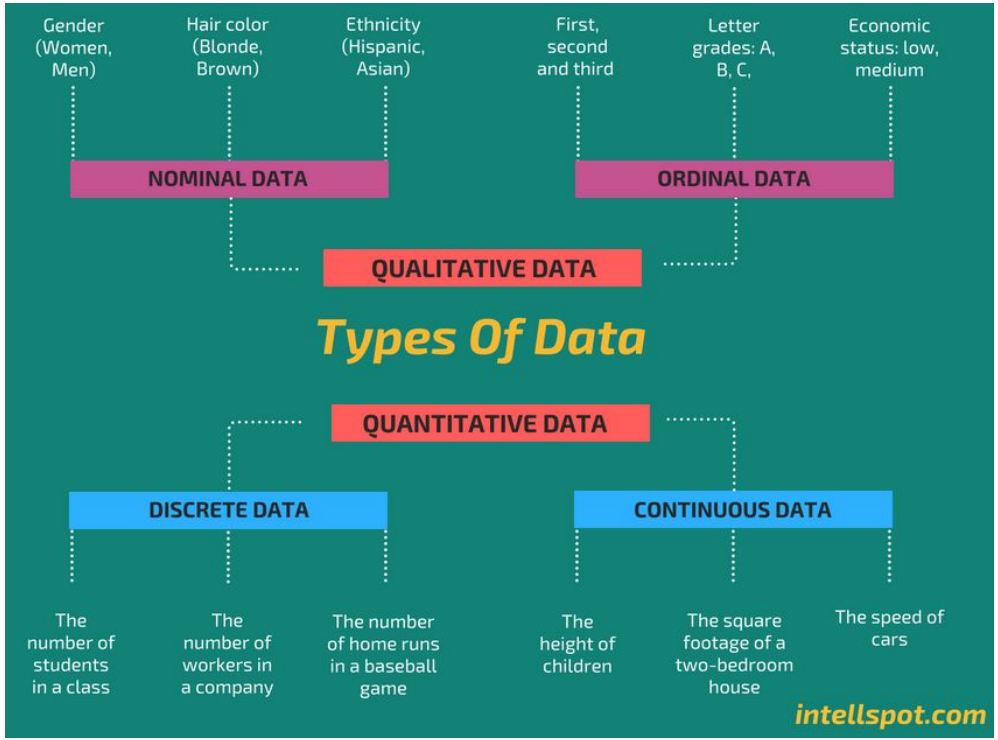
  ⇒ structured data

Qualitative (định tính):

- nominal (định danh)
- binary (định danh True/False)
- ordinal (thứ tự)

  ⇒ unstructured data

(text, phân loại, datetime)

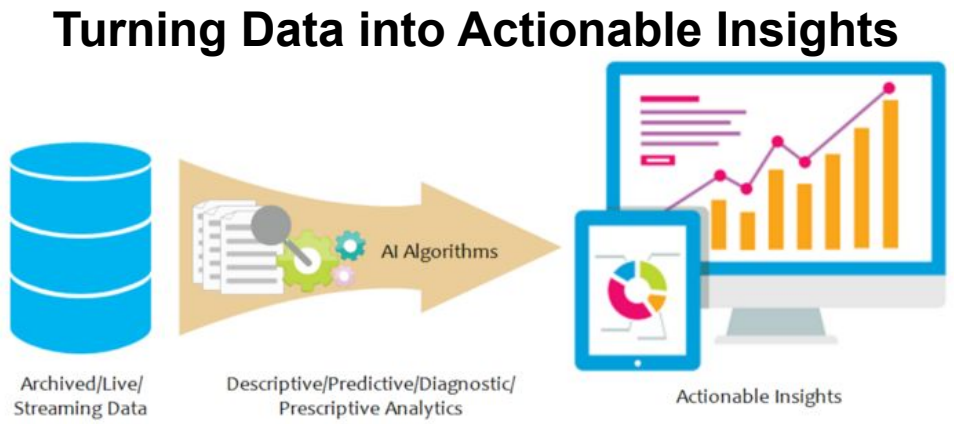


Source: https://www.intellspot.com/data-types/

1.2 Phân tích mô tả là gì?

Các thông tin mà phân tích mô tả cần chỉ ra:

- Các thay đổi trong quá trình hoạt động của tổ chức/doanh nghiệp
- Business performance từ quá khứ đến hiện tại
- Các khuynh hướng/chiều hướng
- Các điểm yếu, điểm mạnh
- Các mối quan hệ



Source: https://towardsdatascience.com/turning-data-into-actionable-insights-c246969fa4c

1.2 Phân tích mô tả là gì?

Các bước của quá trình phân tích mô tả:

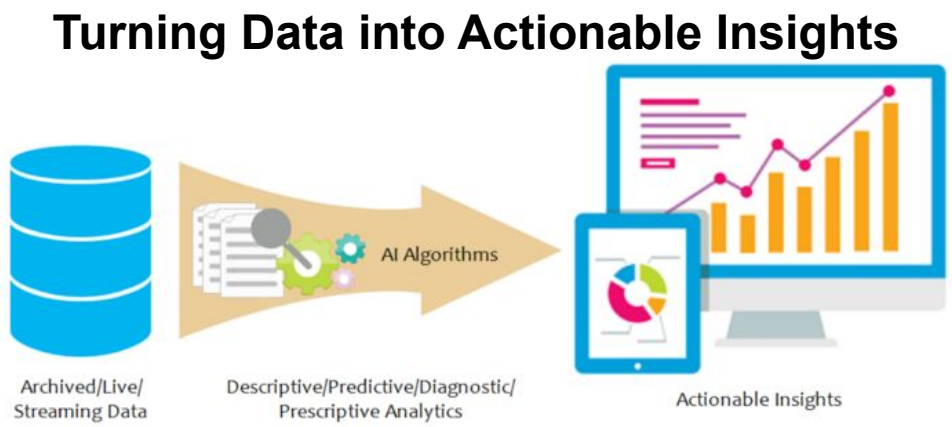
1. Xác định các business metrics	- Chỉ số đo lường cần phân tích (ví dụ: sự thay đổi của giá sản phẩm, doanh thu, lợi nhuận, số người dùng, số người đăng ký… theo năm, quý, tháng, tuần…)
2. Xác định các dữ liệu cần thiết	- Truy vấn các nguồn dữ liệu, từ reports đến databases
3. Chuẩn bị và xử lý dữ liệu	- Làm sạch, chuyển đổi (sang dạng có cấu trúc hoặc dạng phù hợp)
4. Phân tích dữ liệu (analyse)	- Tìm ra các thống kê, khuynh hướng, patterns
5. Trình bày kết quả	- Các biểu đồ - Các tóm tắt

Source: https://studyonline.unsw.edu.au/blog/descriptive-predictive-prescriptive-analytics

1.2 Phân tích mô tả là gì?

Các kết quả của phân tích mô tả:

- Report (báo cáo)/ Survey (khảo sát)
- Các hình minh hoạ
- Các dashboard (bảng điều khiển kỹ thuật số):
  - Thường trình bày một nhóm các KPI chính
  - Các trường thông tin chính
  - Các bảng tóm tắt



Source: https://towardsdatascience.com/turning-data-into-actionable-insights-c246969fa4c

1.2 Phân tích mô tả là gì?

Các loại phân tích mô tả từ cổ điển đến nâng cao:

(Tham khảo: https://uc-r.github.io/descriptive)

1. Phân tích kiểu cổ điển (classical)
   - Phân tích đơn biến (thuộc descriptive statistics)
   - Phân tích đa biến (thuộc descriptive statistics)
   - Phân tích số liệu với ***thống kê suy luận*** (inferential statistics)
2. Khai thác dữ liệu text (text mining)
   - Unstructured information extracting
   - Sentiment analysis
3. Giải thuật học không giám sát (unsupervised learning)
   - Principal Component Analysis
   - Trend analysis
   - Cluster analysis (k-Means, kNNs)

1.2 Phân tích mô tả đơn biến

Các loại phân tích mô tả đơn biến (univariate analysis) cơ bản:

- Measures of frequency (độ đo về tần số):
  Number of Occurrences, Percentage
- Measures of central tendency (độ đo về khuynh hướng tập trung):
  Mean, Median, Mode
- Measures of spread (dispersion/variability) (độ đo về sự mở rộng):
  Range/Quartiles, Variance & Standard Deviation
- Measures of position (độ đo về phân vị):
  Percentiles & Quantiles, Standard Scores
- Measures of shape (độ đo về hình dạng phân bố):
  Skewness/Kurtosis, Normal Distribution

1.2 Phân tích mô tả đa biến

Các loại phân tích mô tả số liệu đa biến (multivariate analysis) cơ bản:

- Covariance (hiệp phương sai)
- Correlation & Coefficient (sự tương quan và hệ số tương quan)

Phần 2.
Một số phân tích mô tả đơn biến cơ bản

2.1. Độ đo về tần số/tần suất
2.2. Độ đo về khuynh hướng tập trung
2.3. Độ đo về sự/tính mở rộng (phân tán)
2.4. Độ đo về phân vị
2.5. Độ đo về dạng phân bố

2.1 Độ đo về tần số/tần suất

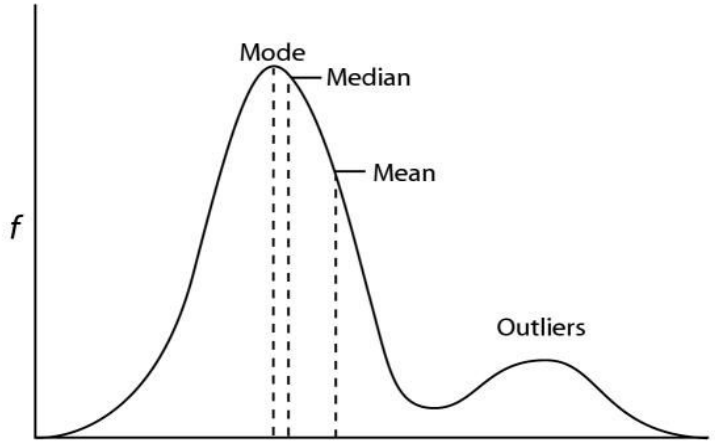
- Number of Occurrences: tần số
- Percentage: tần suất, tỷ lệ phần trăm

```
list.count()
arr.count_nonzero(arr=="...")
np.unique(return_counts=True)
```

2.2 Độ đo về khuynh hướng tập trung

Central tendency: các giá trị mà dữ liệu có xu hướng đạt tới

- Giá trị trung bình (mean): giá trị tiêu biểu giúp tóm tắt dữ liệu
- Giá trị trung vị/giá trị trung tâm (median)
- Giá trị "yếu vị" (mode): giá trị của phần tử có số lần xuất hiện nhiều nhất
- Ngoại lai (outliers): (các) giá trị chênh lệch bất thường so với các giá trị khác



2.2 Độ đo về khuynh hướng tập trung

Giá trị trung bình (mean)

- Population mean: μ **μ**
- Sample mean: x̄
- Lưu ý: có 1 giá trị duy nhất

Population Mean	Sample Mean
$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$
N = number of items in the population	n = number of items in the sample

Diễn giải Luật số lớn (law of large number) theo mean:
Khi kích thước dữ liệu càng lớn, giá trị của sample mean càng gần population mean

```
import numpy as np
mean = np.mean(arr)
```

2.2 Độ đo về khuynh hướng tập trung

Giá trị trung vị/trung tâm (median)

- Là giá trị tách một nửa lớn hơn và một nửa nhỏ hơn của một trường dữ liệu
- Nếu số phần tử là lẻ?
- Nếu số phần tử là chẵn?
- Lưu ý: có một giá trị duy nhất

```
import numpy as np
mean = np.median(arr)
```

2.2 Độ đo về khuynh hướng tập trung

Giá trị yếu vị (mode)

- Một tập có thể có một hay nhiều giá trị mode
- Hoặc cũng có thể không có giá trị mode nào (hoặc tất cả giá trị là mode)
- Lưu ý: các hàm dưới đây so sánh bằng (cẩn thận với số thực!!!)

```
# Tìm mode với np.unique
import numpy as np
values, counts = np.unique(arr, return_counts=True)
mode = values[np.argmax(counts)]
```

```
# Cách khác: dùng module stats trong thư viện scipy
from scipy import stats
mode = stats.mode(arr)
```

```
# Cách khác: dùng thư viện statistics
import statistics
statistics.mode(arr)
```

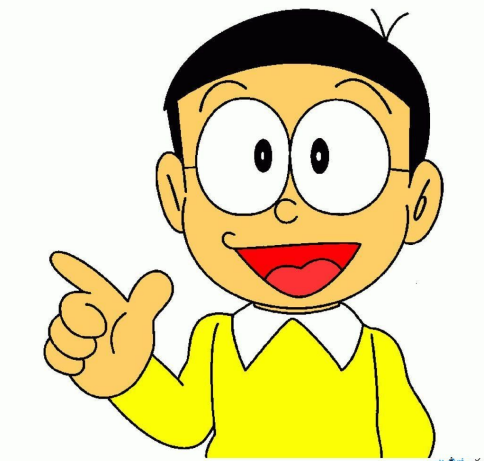
2.2 Độ đo về khuynh hướng tập trung

So sánh tóm tắt giữa Mean - Median - Mode

	Mean	Median	Mode
Có outliers	Bị ảnh hưởng	Ít bị ảnh hưởng	Không bị ảnh hưởng
Ít dữ liệu	Thường không tốt bằng median	Có tính đại diện tốt hơn các giá trị kia	Thường không đại diện cho khuynh hướng tập trung
Dữ liệu lớn và không outlier	Có tính đại diện cao hơn median & mode	Không bằng mean	Không được tạo thành bởi tất cả các giá trị dữ liệu
Về thao tác	Phải gộp hết dữ liệu	Phải gộp hết dữ liệu	Có thể tính mà không cần gộp hết, dễ dàng xác định bằng hình ảnh
Tính duy nhất	Có	Có	Không

So sánh tóm tắt giữa Mean - Median - Mode

Không có giá trị nào là tốt nhất, nhưng việc chỉ sử dụng một giá trị để phân tích chính là cách làm "tệ nhất"



2.3 Độ đo về sự mở rộng

Thuật ngữ: spread, dispersion, variability, scatter

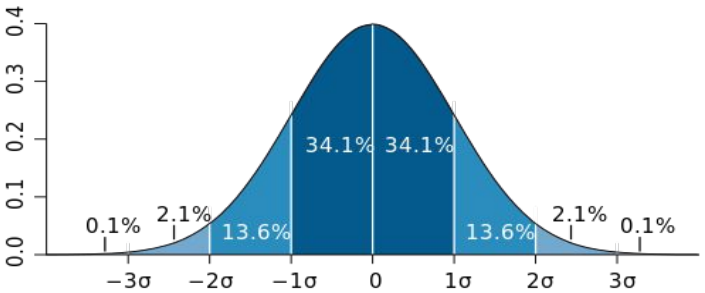
Các độ đo:

- Max/Min
- Range (khoảng)
- Variance (phương sai)
- Standard Deviation (độ lệch chuẩn)

2.3 Độ đo về sự mở rộng

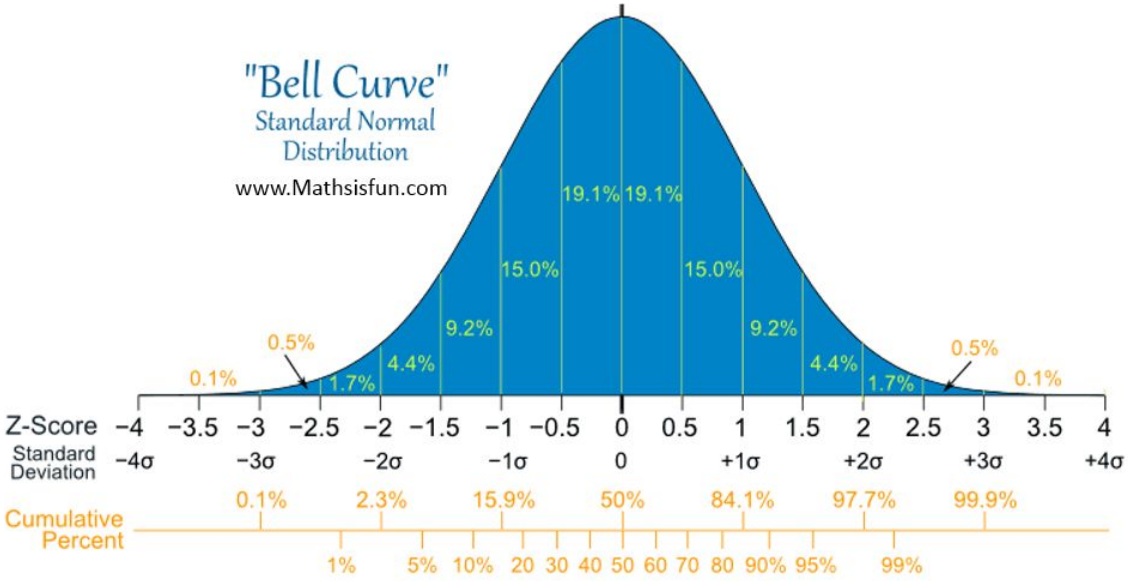
- Variance (phương sai của một biến ngẫu nhiên):
  - Là đại lượng đo tính biến thiên so với mean (measure of variability)
  - Chỉ ra các giá trị của biến đó thường cách mean bao xa
  - Cách tính: trung bình của tổng bình phương độ lệch so với mean
- Standard Deviation (độ lệch chuẩn của một biến ngẫu nhiên):
  - Bằng căn bậc hai của phương sai
  - Đo mức độ phân tán của một tập các giá trị (measure of dispersion)
  - Dễ so sánh với dữ liệu hơn vì cùng đơn vị

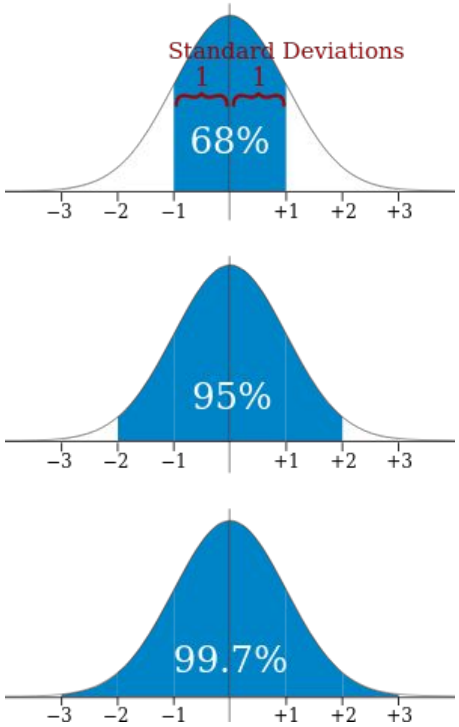
$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2}$$



2.3 Độ đo về sự mở rộng

❖ Variance và Standard Deviation





2.3 Độ đo về sự mở rộng

- (Estimated) Var và (estimated) Std:
  - "Nên/cần" phản ánh mức độ "mở rộng" của dữ liệu quanh population mean
  - Thường không biết population mean nên chỉ có thể ước lượng Var và Std.
  - Lưu ý: do khoảng cách từ data tới sample mean thường có khuynh hướng nhỏ hơn khoảng cách từ data tới population mean
    ⇒ Dễ "đánh giá thấp" Var
    ⇒ Cần chia cho *n - 1* thay vì *n*

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$$ **Population Variance**

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$ **Sample Variance**

Standard Deviation Formula	
Population	**Sample**
$\sigma = \sqrt{\frac{\sum (X - \mu)^2}{N}}$	$s = \sqrt{\frac{\sum (X - \bar{x})^2}{n - 1}}$
X – The Value in the data distribution *μ – The population Mean* *N – Total Number of Observations*	*X – The Value in the data distribution* *$\bar{x}$ – The Sample Mean* *n - Total Number of Observations*

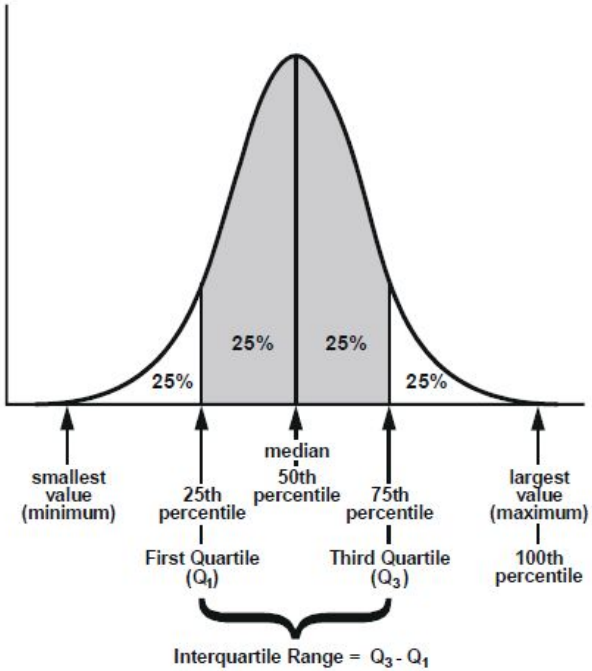
2.3 Độ đo về sự mở rộng

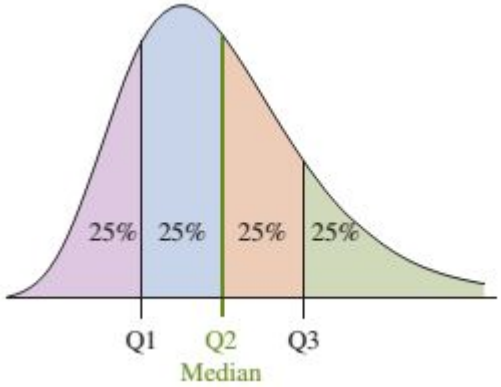
- Max/Min
- Range (khoảng)
- Variance (phương sai)
- Standard Deviation (độ lệch chuẩn)

```
range = max - min =
np.ptp(arr)
var = np.var(arr)
std = np.std(arr)
```

2.4 Độ đo về phân vị

- Percentiles & Quantiles
- Standard Scores

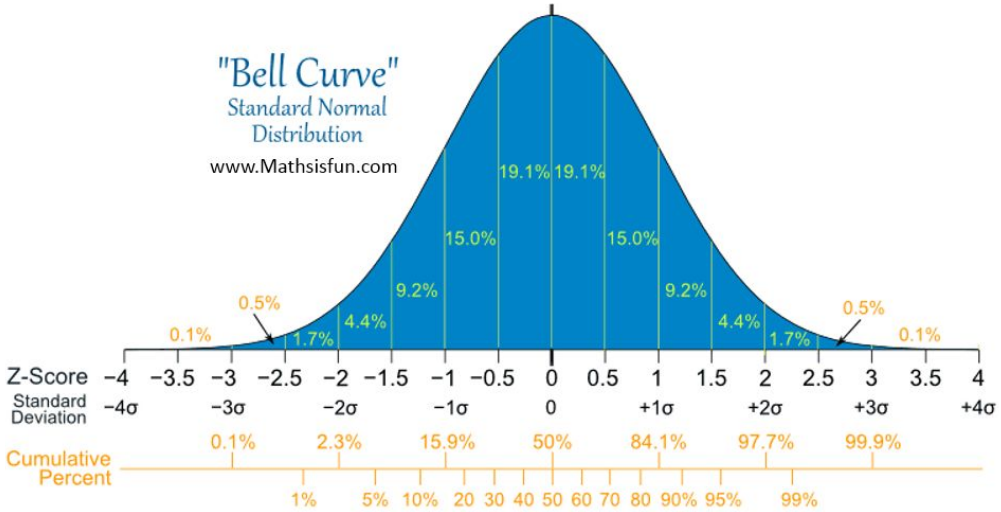




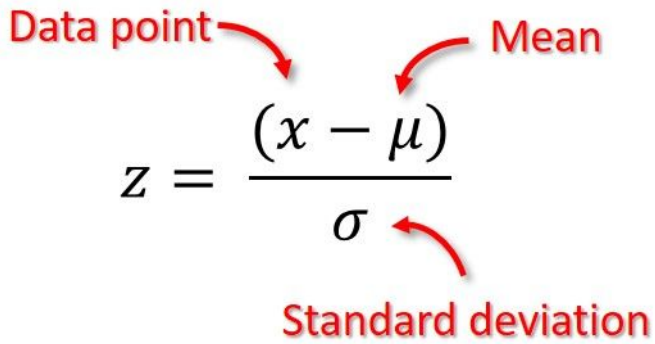
2.4 Độ đo về phân vị

Standard Score hay Z-Score

⇒ Khoảng cách từ một điểm dữ liệu tới mean tính bằng số lần Std



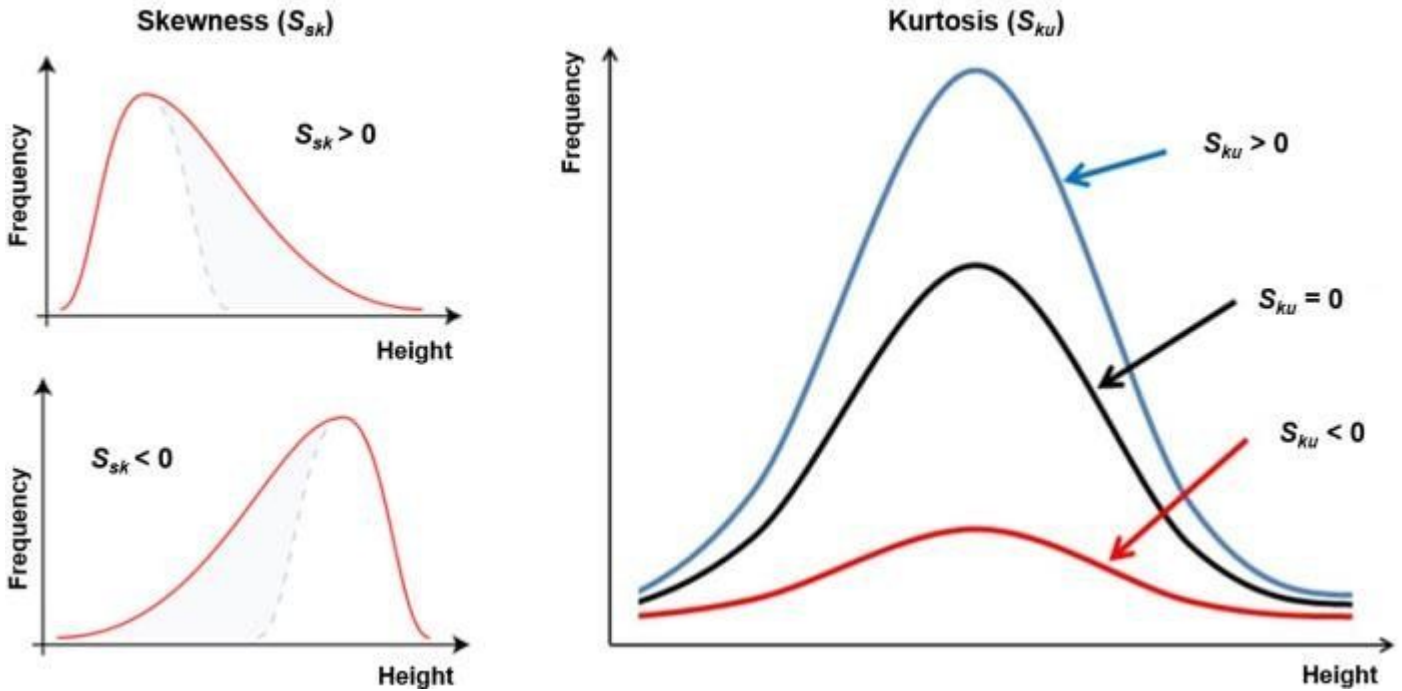
Standard Score hay Z-Score




```
from scipy import stats
stats.zscore(a, axis=None)
# Để tìm zscore của 1 giá trị
stats.zmap(scores=giá-trị, compare=a)
```

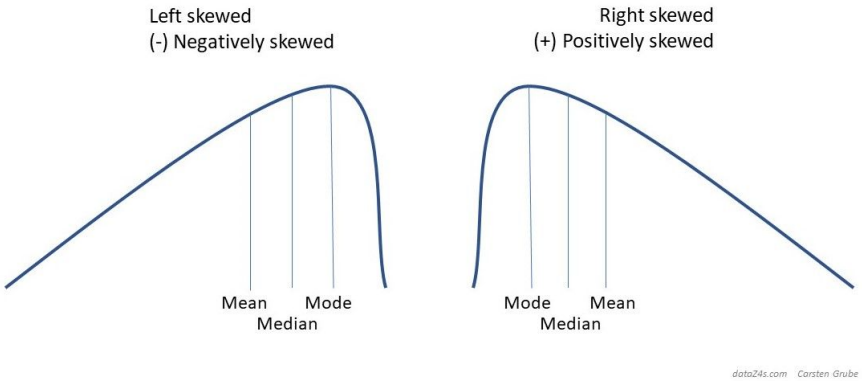
2.5 Độ đo về dạng phân bố

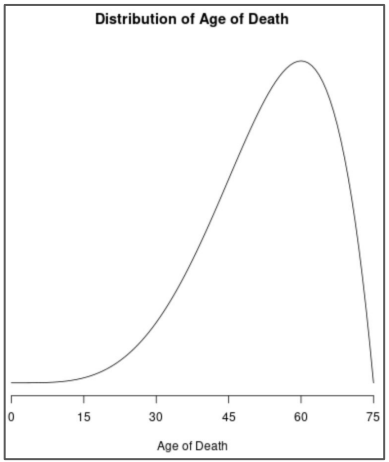
- Skewness/Kurtosis
- Normal Distribution

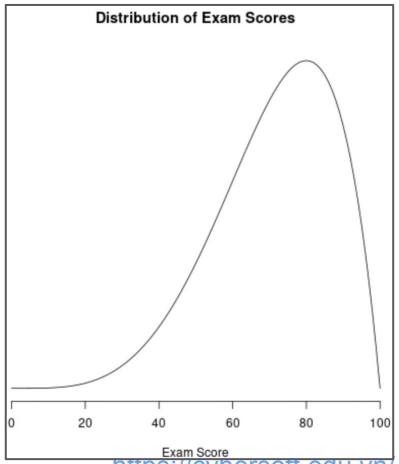


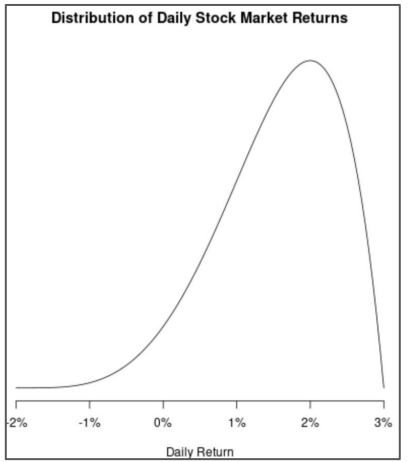
2.5 Độ đo về dạng phân bố

- **Skewness**
  - "Độ lệch"
  - "Chỉ số thiếu đối xứng"





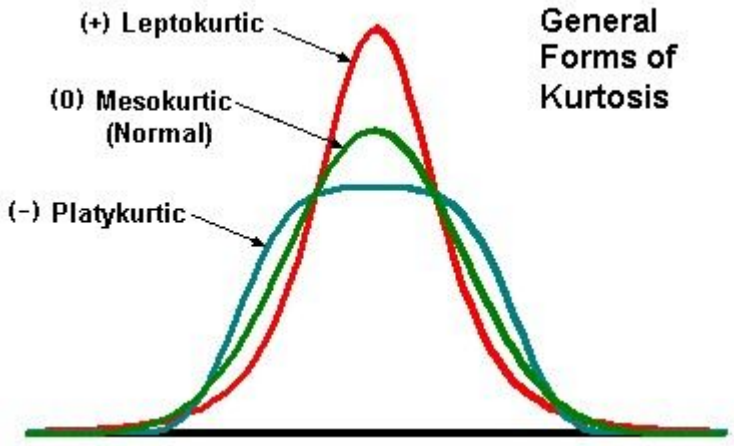




2.5 Độ đo về dạng phân bố

- Kurtosis

  "Độ gù/nhọn"



2.5 Độ đo về dạng phân bố

- Skewness
  (Pearson coefficient)

$$\text{Skewness} = \frac{3(\text{Mean} - \text{Median})}{\text{Standard Deviation}}$$

- Kurtosis

$$\text{Kurtosis} = \frac{\sum \left(x - \bar{x}\right)^4}{(n - 1) \cdot S^4}$$

❖ Hàm tính skewness và kurtosis trong Scipy

```
from scipy import stats
stats.skew((a, axis=0, bias=True, nan_policy='propagate')
stats.kurtosis(a, axis=0, fisher=True, bias=True)
```

Phần 3. Một số phân tích mô tả đa biến cơ bản

3.1. Covariance
3.2. Correlation

Thuật ngữ:

- Bivariate Analysis (phân tích nhị/song biến)
- Multivariate Analysis (phân tích đa biến)

Là loại phân tích thống kê thực hiện trên hai hay nhiều biến phụ thuộc nhau

3.1 Covariance

Hiệp phương sai: mô tả mức độ 2 biến cùng biến đổi với nhau

- Khi x cách xa mean của x thì y cách xa mean của y như thế nào
- Đồng biến (positive covariance)
- Nghịch biến (negative covariance)
- Khoảng giá trị: $-\infty < Cov(x,y) < +\infty$

Population Covariance Formula

$$Cov(x,y) = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{N}$$

Sample Covariance

$$Cov(x,y) = \frac{\Sigma(x_i-\bar{x})(y_i-y)}{N-1}$$

Cách tính covariance với Numpy

```
np.cov(m, y=None, rowvar=True, bias=False, ...)
# m: ma trận có 1 row, hoặc gồm các row cần tính cov
# y: biến cần tính cov với m (nếu muốn tách m ra 2
biến)
# rowvar: set là False nếu muốn tính cov cho cột
# bias: nếu là True sẽ chia cho N, thay vì N-1
```

3.2 Correlation

Sự tương quan (correlation):

- Covariance không xác định được mối quan hệ giữa 2 biến là mạnh hay yếu
- Correlation là giá trị "chuẩn hóa" của covariance, giúp xác định điều này
- Công thức tính hệ số tương quan (correlation coefficient) theo Pearson (R)

$$R = \rho_{xy} = \frac{\mathrm{Cov}(x, y)}{\sigma_x \sigma_y}$$

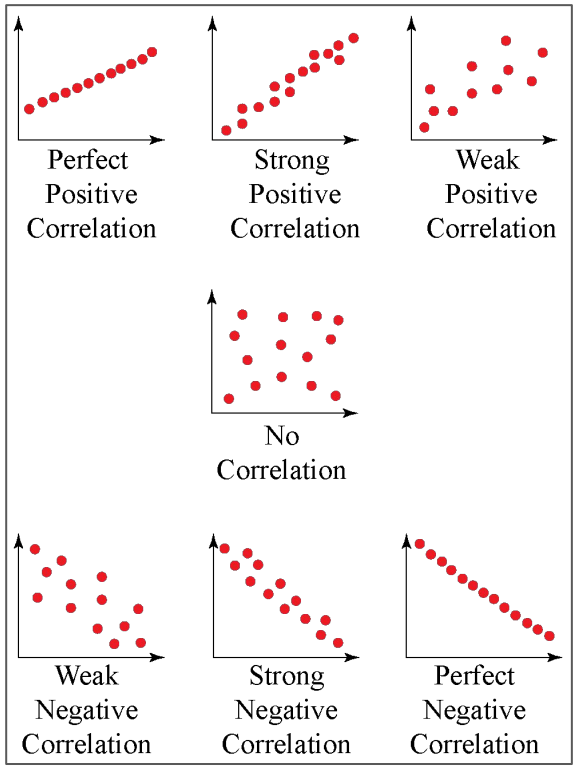
- Correlation thể hiện một tỉ lệ và không có đơn vị
- Khoảng giá trị: -1 < corrcoef < 1

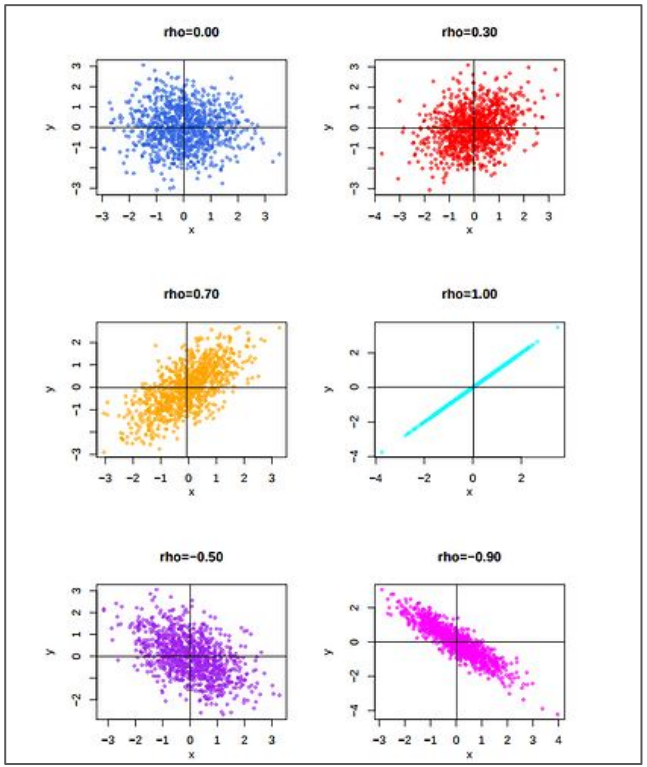
Hệ số tương quan Pearson (correlation coefficient):

mô tả mức độ 2 biến cùng biến đổi với nhau

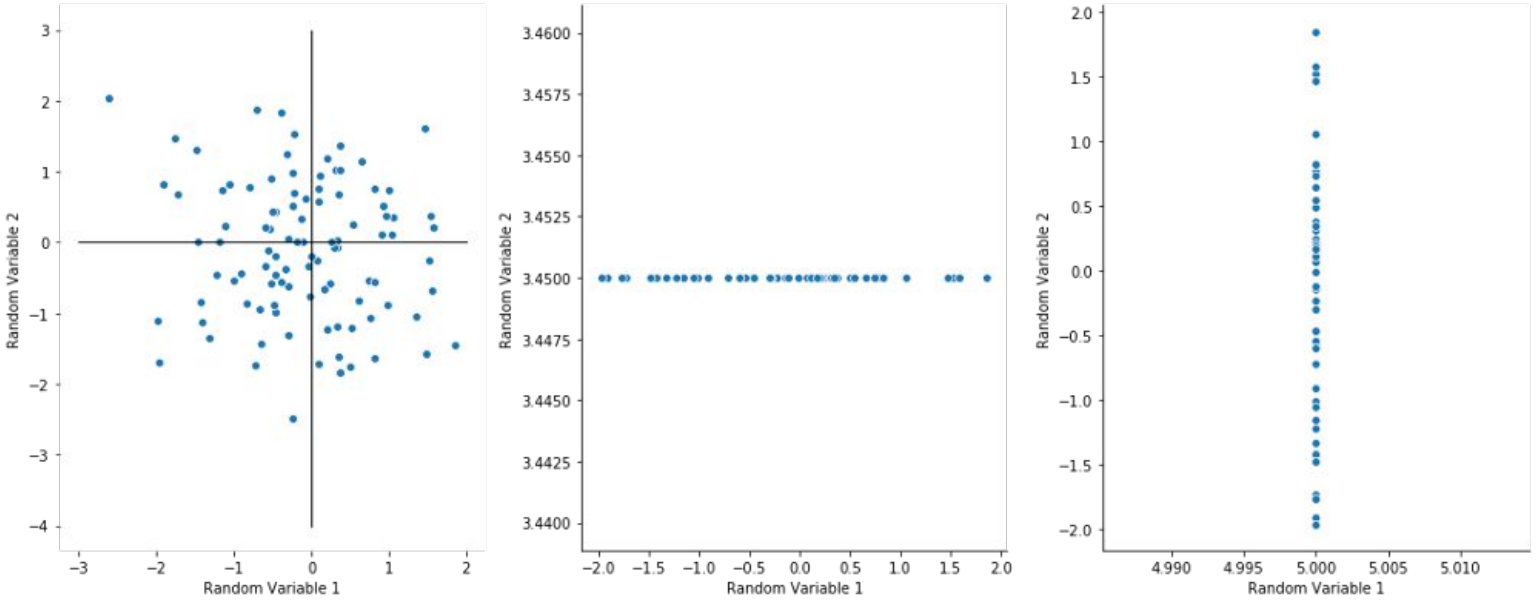
```
np.corrcoef(m, y=None, rowvar=True)
# m: ma trận có 1 row, hoặc gồm các row cần tính cov
# y: biến cần tính cov với m (nếu muốn tách m ra 2
biến)
# rowvar: set là False nếu muốn tính cov cho cột
```

Mức độ của sự tương quan (level of correlation)





Các trường hợp không có sự tương quan



Bài tập - Luyện tập

Bài Tập

1. Thực hiện phân tích mô tả cho đồ án giữa khóa

THANK YOU!